



ADVANCE REGRESSION ASSIGNMENT

Part II



Shishir Khanal

(Msc, ML & AI)

Question 1:

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

Lambda discussed here is indicated by alpha in scikit-learn package.

In ridge and lasso regression, lambda (λ) is a regularization parameter that controls the strength of the regularization applied to the model. Regularization is a technique used to prevent overfitting by adding a penalty term to the loss function, discouraging overly complex models. The optimal value of lambda for our models are mentioned below:

- ❖ Optimal value of lambda for Ridge Regression = **10**
- ❖ Optimal value of lambda for Lasso = **0.001**

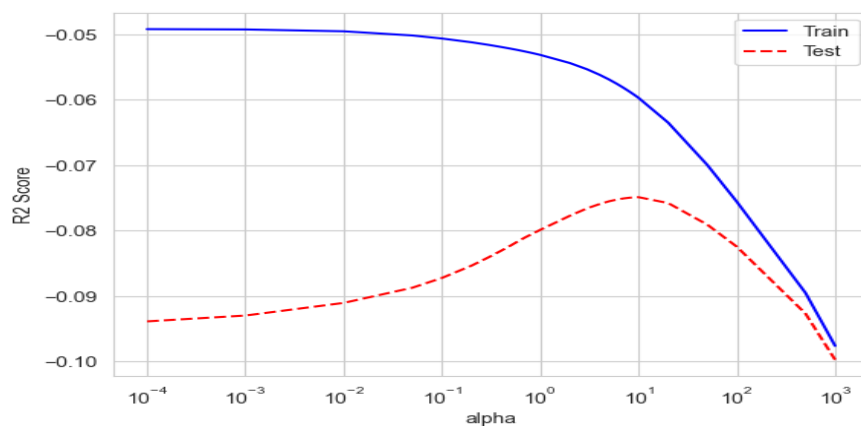


Fig1: Graph representation of lambda for ridge regression

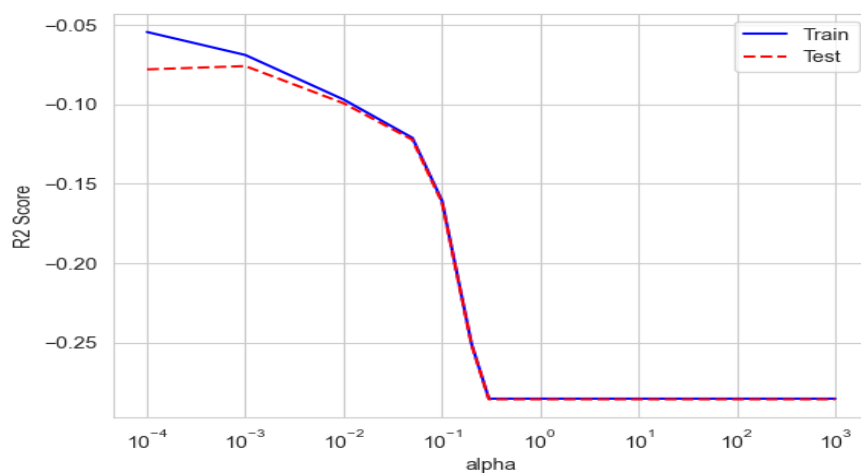


Fig2: Graph representation of lambda for lasso regression

Advance Regression Assignment, Part II

What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

The table below illustrates the metrics of the model with the initial alpha value. i.e.

- ❖ Initial optimal value lambda for Ridge Regression = 10
- ❖ Initial optimal value of lambda for Lasso = 0.001

	Ridge Regression	Lasso Regression
Metric		
R2 Score (Train)	0.95	0.92
R2 Score (Test)	0.91	0.93
RSS (Train)	7.05	11.29
RSS (Test)	3.66	2.92
MSE (Train)	0.01	0.01
MSE (Test)	0.01	0.01
RMSE (Train)	0.08	0.10
RMSE (Test)	0.11	0.10

Fig3: Metrics table of model when initial alpha value

The table below illustrates the metrics of the model with a **doubled alpha** value. i.e.

- ❖ After doubling, value lambda for Ridge Regression = **20**
- ❖ After doubling, value of lambda for Lasso = **0.002**

	Ridge Regression	Lasso Regression
Metric		
R2 Score (Train)	0.93	0.91
R2 Score (Test)	0.93	0.91
RSS (Train)	9.37	13.49
RSS (Test)	2.82	3.45
MSE (Train)	0.01	0.01
MSE (Test)	0.01	0.01
RMSE (Train)	0.09	0.11
RMSE (Test)	0.10	0.11

Fig4: Metrics table of model when double the alpha value

Hence, Changes in Ridge Regression metrics:

- R2 score of train set changed from **0.95 to 0.93**
- R2 score of test set changed from **0.91 to 0.93**

Advance Regression Assignment, Part II

Changes in Lasso metrics:

- R2 score of train set decreased from **0.92 to 0.91**
- R2 score of test set decreased from **0.93 to 0.91**

After doubling the alpha values, the most significant predictor variables are given by:

GrLivArea	1.08
OverallQual_8	1.07
OverallQual_9	1.07
Neighborhood_Crawfor	1.07
Functional_Typ	1.06
Exterior1st_BrkFace	1.06
OverallCond_9	1.06
TotalBsmtSF	1.05
CentralAir_Y	1.05
OverallCond_7	1.04
Name: Ridge, dtype: float64	

Fig 5: Ridge variables

GrLivArea	1.11
OverallQual_8	1.09
OverallQual_9	1.08
Functional_Typ	1.07
Neighborhood_Crawfor	1.07
TotalBsmtSF	1.05
Exterior1st_BrkFace	1.05
CentralAir_Y	1.04
YearRemodAdd	1.04
Condition1_Norm	1.03
Name: Lasso, dtype: float64	

Fig6: Lasso variables

Hence, after doubling the alpha values, the most significant (Lasso predictor) variables are:

- GrLivArea
- OverallQual_8
- OverallQual_9
- Functional_Typ
- Neighborhood_Crawfor
- TotalBsmtSF
- Exterior1st_BrkFace
- YearRemodAdd
- Condition1_Norm

Question 2:

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer:

The choice between Ridge and Lasso regression depends on the specific characteristics of our data and the goals of our analysis. Both Ridge and Lasso are regularization techniques used to address the issue of multicollinearity and overfitting in linear regression models by adding a penalty term to the objective function.

Advance Regression Assignment, Part II

However, they have some differences in terms of how they handle this regularization.

Generally, if we have too many variables and one of our primary goal is feature selection, then we will use Lasso.

If we don't want to get too large coefficients and reduction of coefficient magnitude is one of our prime goals, then we will use Ridge Regression.

The table illustrates the difference in metrics between ridge and lasso regression for our model.

	Ridge Regression	Lasso Regression
Metric		
R2 Score (Train)	0.95	0.92
R2 Score (Test)	0.91	0.93
RSS (Train)	7.05	11.29
RSS (Test)	3.66	2.92
MSE (Train)	0.01	0.01
MSE (Test)	0.01	0.01
RMSE (Train)	0.08	0.10
RMSE (Test)	0.11	0.10

Fig7: Metrics table of ridge and lasso regression

In our models, **I select lasso regression** for the final model prediction as its R2 score is slightly higher, and the gap between training and testing scores is lower. Additionally, lasso regression offers a simpler model with fewer variables compared to ridge regression.

Question 3:

After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

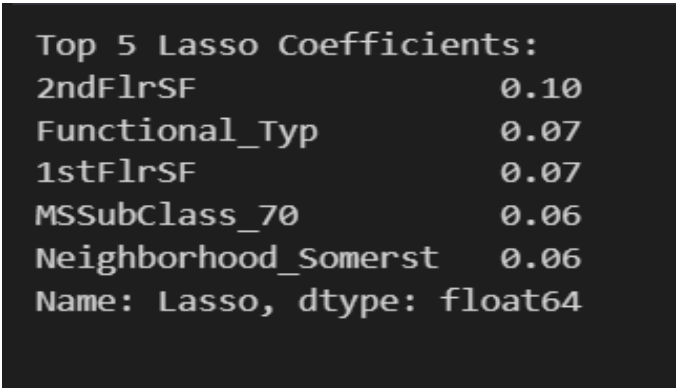
Answer:

Here, we will drop these top 5 features identified by the Lasso model to build a new model.

- OverallQual_9
- GrLivArea
- OverallQual_8
- Neighborhood_Crawfor
- Exterior1st_BrkFace

After dropping our top 5 lasso predictors, we get the following new top 5 predictors:

- 2ndFlrSF
- Functional_Typ
- 1stFlrSF
- MSSubClass_70
- Neighborhood_Somerst



```
Top 5 Lasso Coefficients:
2ndFlrSF          0.10
Functional_Typ    0.07
1stFlrSF          0.07
MSSubClass_70     0.06
Neighborhood_Somerst 0.06
Name: Lasso, dtype: float64
```

Fig8: New top five lasso predictors

Question 4:

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer:

Ensuring that a machine learning model is robust and generalizable is important for its performance in real-world scenarios. Robustness refers to the model's ability to handle variations in the input data, while generalizability refers to its capacity to perform well on unseen data.

A model is considered robust when its performance remains relatively unaffected by variations in the data.

A generalizable model demonstrates the ability to adapt effectively to new, previously unseen data originating from the same distribution used to create the model.

To ensure both robustness and generalizability, it is essential to prevent the model from overfitting. Overfitting occurs when a model exhibits excessive variance, making it highly sensitive to even the slightest alterations in the data. Such a model will identify all the patterns of a training data but fail to pick up the patterns in unseen test data. The model should not be too complex in order to be robust and generalizable.

A robust and generalizable model is likely to have better accuracy on new, unseen data. Overfitting may lead to high accuracy on training data but poor performance on new data. It is important to strike a balance between model accuracy and complexity. Regularization helps balance model complexity, improving accuracy on new data.

The End!!